DOCUMENT RESUME

ED 289 888 TM 870 650

AUTHOR Lynch, Kathleen Bodisch
TITLE Practices in Educational Program Evaluation, 1980-1983.
PUB DATE Apr 87
NOTE 40p.; Paper presented at the Annual Meeting of the American Educational Research Association (Washington, DC, April 20-24, 1987).
PUB TYPE Speeches/Conference Papers (150) -- Reports - Descriptive (141)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Educational Practices; *Effect Size; Evaluation Criteria; Evaluation Methods; Federal Government; National Surveys; Program Content; *Program Effectiveness; Program Evaluation; *Program Proposals; *Program Validation; Success; *Validated Programs
IDENTIFIERS Department of Education; *Joint Dissemination Review Panel

ABSTRACT

Current practice in educational program evaluation was examined through analyses of 232 reports submitted, from 1980 to 1983, by institutions seeking approval for their programs from the U.S. Department of Education's Joint Dissemination Review Panel (JDRP). The JDRP reviews these reports to determine whether educational programs have demonstrated that they are effective. This study also examined how evaluation methods differed for programs which were approved or not approved by the JDRP during this time period. Certain features were found more often in programs approved than those not approved. These included: (1) the presence of an independent evaluator affiliated with a research firm; (2) the use of more than one evaluation design; (3) the absence of obvious errors in data analyses; and (4) the implementation of evaluation designs of high quality. Features of the educational programs and their evaluations were documented through content analyses. Descriptive profiles were developed for the entire sample, as well as the subsamples of approved and not-approved programs. Regression analyses were used to relate differences in evaluation methodology to differences in the programs' effect size. The appended tables include summary data on program content, approvals, grade level, evaluators, outcome measures used, test validity and reliability, evaluation design, data analysis, and effect size. A 36-item reference list concludes the document. (MAC)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

BEST COPY AVAILABLE

Practices in Educational Program Evaluation, 1980-1983

Kathleen Bodisch Lynch, Ph.D.
Office of Medical Education
University of Virginia School of Medicine
Box 382
Charlottesville, Virginia 22908

ED289888

A paper presented at the
American Educational Research Association Annual Meeting
Washington, D.C.
April 1987

Abstract

Current practice in educational program evaluation was examined through analyses of reports submitted by educational programs seeking approval from the U. S. Department of Education's Joint Dissemination Review Panel (JDRP). The JDRP reviews these reports to determine whether educational programs have convincingly demonstrated that they are effective. In this study, features of the educational programs and their evaluations were documented through content analyses of 232 reports submitted to the JDRP from 1980 through 1983. Descriptive profiles were developed for the sample as a whole, as well as for the subsamples of approved and not-approved programs. Regression analyses were used to relate differences in evaluation methodology to differences in the size of educational effects detected by the programs.

TM 870 650

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☒ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

K. B Lynch

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Practices in Educational Program Evaluation, 1980-1983

Kathleen Bodisch Lynch, Ph.D.

Office of Medical Education
University of Virginia School of Medicine
Box 382
Charlottesville, Virginia 22908
804-924-2563

A paper presented at the
American Educational Research Association Annual Meeting
Washington, D. C.
April 1987

Practices in Educational Program Evaluation, 1980-1983

The birth of educational program evaluation as a distinct field of study in the United States may be traced to the Elementary and Secondary Education Act (ESEA) of 1965. Included in this legislation, which mandated a number of federally funded programs to improve the performance of disadvantaged children, was the requirement that educational projects be accountable for their use of federal monies. Congress wanted to know whether large expenditures of funds helped to bring about any real improvements in education.

Acting on the federal requirements, educators began to evaluate their own efforts. Unfortunately, their task was made complicated both by the absence of clear guidelines from Congress as to what questions needed to be answered, and by the fact that program evaluation was not what educators were typically trained to do. These conditions set the stage for scholars to turn their attention to the development of theories and models of educational program evaluation. By 1973, the field had become well enough established that Worthen and Sanders were able to assemble a book about the varieties of evaluation approaches, with chapters contributed by many well-known theoreticians and practitioners in education, psychology, and other areas of social science research.

The proliferation of approaches to educational program evaluation spawned by the 1965 Congressional mandate resulted in a further request by Congress, in the Education Amendments of 1978. Now, a study of the educational evaluation practices themselves was desired. In response, the National Academy of Science, at the invitation of the (then) Office of Education, undertook a study whose purpose was to recommend ways of increasing the effectiveness and usefulness of the Office of Education's evaluation efforts. This study, the results of which were compiled in a 1981 book edited by Raizen and Rossi, is one of a very few systematic reviews of the types and quality of educational evaluations. A more recent effort describing "the state of the art and the sorry state of the science" of evaluation was reported by Lipsey, Crosse, Dunkle, Pollard, and Stobart (1985). The need for documentation of evaluation practice in real-life settings has long been noted by respected researchers and practitioners (see, e.g., Boruch & Cordray, 1980; Cook & Gruder, 1978; Smith, 1979). The improvement of educational program evaluation depends on the accumulation of knowledge about the successes and failures of evaluation methods as actually applied in the classroom.

Background and Purpose of the Study

The primary objective of this study was to document current practice in educational program evaluation through a systematic analysis of reports submitted to the U. S. Department of Education's Joint Dissemination Review Panel (JDRP) during a four-year period. A secondary purpose was to determine how evaluation methods differed for programs which were approved or not approved by the JDRP during this time period.

The JDRP is a body of experts from the Department of Education whose purpose is to review educational products and practices from all over the United States in order to determine whether they are effective. The JDRP makes its judgments through consideration of 10-page written reports supplied by the programs seeking JDRP approval. These reports describe the program's goals, activities, costs, implementation requirements, evaluation procedures, and evidence of effectiveness. The JDRP reviews the reports to determine whether educational programs have convincingly demonstrated that they are effective. Program staff also make an oral presentation before a subgroup of three to seven members of the JDRP. Decisions concerning approval or rejection of an educational program are made on the basis of a simple majority vote of this subgroup.

The JDRP considers educational products and practices which it has approved to be worthy of nationwide dissemination. Consequently, approval is given to a program only if two major conditions are met: (a) the program must persuasively demonstrate that it is effective, and (b) the program must be effective to an exemplary degree. To satisfy the first condition, programs must document their evaluation procedures and show that the observed effects are attributable to the program itself, rather than to other plausible explanatory factors (Datta, 1977; Tallmadge, 1977). In addition, the intervention and its effects should be replicable.

To satisfy the second condition (i.e., that the degree of program effectiveness is exemplary), the program must convince the JDRP that the effects produced are of sufficient magnitude to be considered statistically and educationally significant. While statistical significance can be assessed through conventional techniques of data analysis, the determination of what constitutes educational significance is more difficult. The JDRP uses a combination of judgmental and normative approaches (Sechrest & Yeaton, 1981) for evaluating the educational significance of the size of effects produced by the programs it reviews. Panel members, as experts in the field, render professional judgments about the practical significance of the observed effects. They also try to determine whether the program has convincingly shown that the gains produced exceed what is typically reported in the literature.

Method

During the years 1980 through 1983, 240 educational programs applied to the JDRP. Of these, 232 specified changes in knowledge, attitude, or behavior in students, teachers, parents, paraprofessionals, or other individuals. The other eight programs targeted institutional change, and were excluded from further analyses.

A coding form was developed by the author in order to extract information from the reports submitted to the JDRP. The items and format chosen were based on a review of the literature on assessment of the adequacy of research or evaluation reports, and the literature on meta-analysis (see, e.g., Bernstein & Freeman, 1975; Fang, 1981; Glass, McGaw, & Smith, 1981; Gordon & Morse, 1975; Lipsey, 1983; Sanders & Nafziger, 1976). Data were collected on characteristics of the educational programs described in the JDRP submittals, the methods used to evaluate them, the size of effects produced, and the quality of the written report. This paper will focus on the evaluation designs, and on the relationship of different evaluation characteristics to both effect size and JDRP approval.

The items developed to document the evaluation practice represented by the educational programs applying to the JDRP during the years 1980 through 1983 were organized into three sections: (a) descriptions of the procedures used to measure program outcomes, (b) descriptions of the evaluation designs, and (c) descriptions of the methods of data analysis. Each of these will be discussed. In addition, the procedures used to calculate effect sizes and to relate differences in effect size to differences in evaluation characteristics will be described.

Measurement Features

The affiliation of the evaluator of the educational product or practice for which claims of effectiveness were presented to the JDRP was noted--whether the evaluator was on the program staff or was an external consultant. Information was also gathered to describe the tests or other instruments used to measure the effects of implementing the educational product or practice. Included were items on the derivation of the test--whether published, developed specifically for the project, or adapted from another test; the types of validity and reliability data reported in the submittal; and the appropriateness of procedures followed during test administration. With regard to the last item, in the case of norm-referenced tests, the date on which the tests were administered was noted; in order for these test scores to be interpretable, the tests should have been administered close to the time of year at which the empirical norms were established (usually, not more than two weeks on either side of the norming date). For outcome measures involving treatment and comparison or

control groups, tests should have been administered to both groups at or near the same time.

Design Features

In order to identify the various types of evaluation designs reported in the JDRP submittals, a series of descriptive phrases was listed on the coding form. Any combination of these phrases could be recorded to describe a particular evaluation design. Response categories included such characteristics as the number of groups involved, whether the groups represented no-treatment or alternate-treatment conditions, whether there was pre- or posttesting, whether assignment to groups was randomized or not, and whether norms were used for comparison purposes. This descriptive approach was used in place of assigning labels to the evaluation designs in order to document as completely and objectively as possible the evaluation practice represented in the submittals, without having to force innovative or patched-together designs into predefined categories.

The internal validity of each design identified was rated on a scale ranging from very low to very high. Although standards for judging the quality of research designs are by no means uniformly agreed upon (Hirschi & Selvin, 1967; McTavish, Brent, Cleary, & Knudsen, 1975), Campbell and Stanley (1963) and Cook and Campbell (1979) have outlined threats to validity typically associated with various approaches to conducting research and evaluation studies. Their conceptualization is generally well respected in the field, and was used as a reference point in assigning the ratings of design quality. For each design, a judgment was made concerning the degree to which threats to internal validity could be discounted. A rating of very high was given when all of the applicable threats to internal validity could be ruled out, enabling the reviewer to conclude with a reasonable degree of certainty that it was the educational program seeking JDRP approval which produced the claimed effects. A rating of very low was given to designs in which there was a "fatal flaw"--that is, where at least one of the threats to internal validity could be considered a compelling and plausible rival explanation for the results obtained. Ratings between very high and very low were assigned as follows. Evaluation designs were rated high when all but one or two of the threats to internal validity could definitely be ruled out, when neither of the possible threats could be considered a fatal flaw, and when, overall, the evidence that the program caused the observed effects was believable. A medium rating was assigned when at least half of the threats could be ruled out, there were no fatal flaws, and, overall, the evidence was ambiguous--neither totally convincing nor totally unconvincing. Designs were rated low when fewer than half of the threats to internal validity could be ruled out, and he evidence was not very convincing, but there were no fatal flaws.

Finally, data on the external validity of the evaluation designs were collected. Any evidence provided in a submittal to indicate that a program had been successfully replicated was recorded on the coding form.

Data Analysis Features

In this section, the descriptive and inferential statistical analyses applied to the outcome data were documented. Whether the analysis techniques were appropriate for the type of data gathered was also noted. Additionally, the form in which test results were reported in the JDRP submittals was recorded, as well as whether tests of statistical significance were used to identify differences between groups.

Calculation of Effect Sizes

Effect sizes were obtained by transforming program results into standard scores for all programs for which the necessary data were supplied. In the simplest cases, for studies involving a comparison between a treatment and a no-treatment group, the difference between the means of the treatment group and the comparison group was divided by the standard deviation of the comparison group (Glass, 1977). This allows one to describe the status of the treatment group by reference to the distribution of outcome scores which would have been expected in the absence of any intervention. In cases where results were based on other evaluation designs, adaptations of the basic effect size formula were used, following recommendations reported in the literature (see, for example, Bryant, 1982; Glass, 1980; McGaw & Glass, 1980).

After effect sizes were calculated, they were related to characteristics of the educational programs and evaluations through traditional methods of data analysis. In this paper, only the relationships to evaluation characteristics will be discussed. Mean effect sizes were calculated for different levels of categorical variables, Pearson product-moment correlations were obtained between effect sizes and continuous variables, and analyses of variance and linear regression were used to identify the proportion of variance in the distribution of effect sizes which was accounted for by characteristics of the evaluations.

During the four years covered by this study, 165 out of 232 program narratives reviewed by the JDRP (or 71%) provided the data necessary to calculate effect sizes. Some JDRP reports provided effectiveness data for more than one content area, target audience, type of objective, type of outcome measure, type of evaluation design, and grade level. Effect sizes were computed separately for each of these variables for each program. Within programs, effect sizes were then aggregated across grade level. When more than one outcome measure was used for a program, effect

sizes were also aggregated across tests within each of two types of outcome measures--published and locally developed. Consequently, the number of effect sizes retrieved from each report varied, with a total of 263 effect sizes obtained.

Results

Description of the Educational Programs

The author summarized data from 232 reports submitted to the JDRP from January 1, 1980 through December 31, 1983. The number of submittals reviewed in each of these years ranged from 45 to 68. Sixty-two percent of all the submittals were approved, with percentages by year varying from 57% to 69%, as can be seen in Table 1.

Insert Table 1 about here

A wide variety of content areas were addressed in the submittals, and some reported on educational programs in more than one area; a total of 326 programs were described. Over 50% of the programs had objectives related to reading or math. The next most frequent content areas were, in order, special education, career education, language arts, natural science, social science, and health/physical education--39% of the programs could be classified into these categories. Table 2 presents a complete listing of the content areas addressed in the submittals.

Insert Table 2 about here

In almost all cases (n = 304) the target audience for the programs was students, ranging from preschool through graduate school. For a few programs, the target audience was teachers or administrators, adult learners, or parents. Table 3 presents the distribution of programs across the various grade levels. As might be expected, most of the programs were developed for school-aged children (K through grade 12), with more efforts occurring in the elementary and middle schools (K through grade 6) than in the junior and senior high schools (grades 7 through 12).

Insert Table 3 about here

The types of objectives that the educational programs addressed were categorized as being either cognitive, behavioral, or attitudinal/affective. Almost every submittal (n = 224, or 97%) presented at least one objective in the cognitive domain, with behavioral and attitudinal objectives occurring much less frequently (21% and 16% of all submittals, respectively). While very few programs were designed to effect only behavioral (n = 4) or only attitudinal (n = 2) changes, programs with only cognitive objectives were common (n = 157, or 68% of all submittals). The frequencies presented in Table 4 show the number of times the different types of objectives and combinations of objectives were addressed in the submittals reviewed for this study.

Insert Table 4 about here

Description of the Program Evaluations

Evaluators. Although the JDRP does not require submittals to identify the evaluators of their programs, over 75% of the submittals (n = 176) reviewed during the time period of this study provided such information. For a small group of the submittals, program staff had the sole responsibility for program evaluation efforts. In over half of the cases, however, independent evaluators conducted the evaluations, either alone or in combination with program staff or other types of evaluators, such as program developers or representatives from district research and evaluation offices. Of the independent evaluators, most were identified as having academic affiliations, the rest being associated with research or consulting firms. Table 5 displays the data concerning evaluators' affiliations retrieved from the JDRP submittals.

Insert Table 5 about here

Outcome measures. Data were collected on all instruments which measured outcomes for which claims of effectiveness were made. The three types of objectives--cognitive, behavioral, and attitudinal/affective--were included. While in most cases (n = 113, or in 49% of the submittals) programs based their claims of effectiveness on data from a single outcome measure, the number of instruments described in each submittal varied from 0 to 7. In some submittals, only one instrument was used to measure all program outcomes (e.g., reading and math scores from the same standardized test). In other submittals, more than one instrument

was used to measure a single program outcome (e.g., scores from both published and locally developed tests of math achievement). A total of 438 outcome measures were described in the 232 JDRP submittals.

The type of instruments most frequently chosen to measure program outcomes (n = 250, or 57% of all instruments reported) were published tests such as the MAT (Metropolitan Achievement Tests), CAT (California Achievement Tests), and SAT (Scholastic Aptitude Tests). Twenty-nine percent of the instruments used (n = 129) were locally developed outcome measures created for specific programs. In a few cases, instruments were adopted from other sources but modified to make them more relevant to the program being evaluated. Concerning test administration, in evaluations involving treatment and comparison groups, in 99% of the cases tests were given to both groups at the same time; in evaluations based on a norm-referenced design, tests were administered at the appropriate norming times in only 70% of the cases. Table 6 presents data about these and other features of the outcome measures described in the JDRP submittals.

Insert Table 6 about here

The amount of information which was provided about the validity and reliability of the outcome measures varied. For the majority of the instruments, at least one type of validity and one type of reliability was reported (63% and 62%, respectively). In a much smaller percentage of cases, more than one type of validity or reliability were cited. The type of validity most frequently reported was content validity (53% of the instruments), and the type of reliability most frequently reported was internal consistency (38%). It should be noted that in many cases, the only statement made with regard to the validity or reliability of the instrument was the "the manual stated that these were high." Moreover, for about one-fourth of the instruments, no information at all was presented as to their validity or reliability.

Evaluation designs. Data were collected on each evaluation design used to gather evidence to substantiate the claims of effectiveness made in each submittal. Recall that a submittal could describe one or more educational programs, and each program could have one or more types of objectives. Each of these objectives, in turn, could be measured by one or more instruments, and each of these instruments could have been administered in accordance with the requirements of a different evaluation design. For example, in many cases a norm-referenced evaluation design was used with a published achievement test and a nonequivalent control group design was used with a locally developed test, in order to

provide complementary evidence that a particular objective had been achieved. In the 232 JDRP submittals reviewed, a total of 363 evaluation designs were reported.

The types of evaluation designs described in the submittals were documented on the coding forms by using combinations of descriptive phrases. Table 7 presents the frequencies of occurrence of the designs based on the coding form categories.

Insert Table 7 about here

The evaluation design most frequently employed by the programs was the nonrandomized pre-post comparison group design, referred to as the nonequivalent control group design by Campbell and Stanley (1963). A total of 87 submittals (or 38%) reported using this design. Next most frequently used (n = 67, or 29%) was the norm-referenced design, which involves pre-post comparisons of scores based on published norms (Tallmadge & Wood, 1978). Following this in frequency was the nonrandomized post-only comparison group design (n = 57, or 25%)--the nonequivalent control group design without the pretest scores. The randomized pre-post control design (one of Campbell and Stanley's "true experimental" designs) and the one-group pre-post design (a "pre-experimental" design) occurred with almost equal frequency (n = 30 or 13%, and n = 33 or 14%, respectively) while all other types of designs occurred fewer than 10 times each.

When categories of designs from the coding form with similar characteristics were aggregated, the frequencies presented in Table 8 resulted. Quasi-experimental designs, which include the

---

Insert Table 8 about here

---

nonequivalent control group, simple time series, and other designs all characterized by nonrandomized assignment of subjects to treatment or comparison groups, were by far the designs most frequently reported in the submittals (n = 170, or 73%). Next in frequency were the norm-referenced designs (n = 88, or 38%). True experimental (randomized assignment to groups) and pre-experimental (one-group, non-time series) designs were found equally in the submittals (n = 50, or 22% each) and only one qualitative design was reported in support of claims of effectiveness. The type of evaluation design used appeared to have little influence on whether or not an educational program

received JDRP approval; as can be seen in Table 8, the different types of evaluation designs had similar rates of approval.

The quality of each evaluation design was rated on a scale from very low to very high. The results are presented in Table 9.

Insert Table 9 about here

The ratings which were assigned tended to cluster in the medium and high categories, which represented situations where the evidence of effectiveness was ambiguous (medium quality) or reasonably convincing (high quality). Overall, more than half of the evaluation designs (i.e., 56%) were rated as medium or worse, indicating a failure to produce reasonably believable evidence that the educational program was responsible for producing the observed changes in the groups receiving the educational product or practice. Ratings indicating certainty that effectiveness was not demonstrated (very low) exceeded those indicating certainty was demonstrated (very high) by a ratio of over 2:1. However, when frequencies of ratings were aggregated across the categories of very low and low, and very high and high, this ratio reverses itself; many more evaluations provided convincing evidence of effectiveness (n = 159, or 44%) than provided convincing evidence that the program was not effective (n = 69, or 19%). JDRP approval was given a greater proportion of the time to submittals with evaluation designs rated as high or very high (79%) than to those rated either medium (69%) or low and very low (32%).

The distribution of the quality ratings was also broken down according to categories of designs in order to identify differences in rated quality dependent on design type. The true experimental designs received greater proportions of very and high ratings than any of the other designs, and these proportions far exceeded those for the sample as a whole. Conversely, true experimental designs had smaller proportions of low or very low ratings than any of the other designs, as well as the overall sample. Quasi-experimental designs received the next best distribution of ratings, followed by the norm-referenced designs. One-group designs had the poorest ratings, receiving disproportionate amounts on both the high and low ends of the quality continuum. Table 10 presents summary data on design type and quality ratings which illustrate these relationships. Mean

---

Insert Table 10 about here

---

ratings for each category of design are also presented, not so much because they have an inherent significance, but rather because they help to convey the order with which the different types of evaluation designs were arrayed along the quality dimension (when quality is defined as the extent to which threats to internal validity can be ruled out).

All submittals provided some evidence of the external validity of their programs by presenting data indicating that the intervention was effective for more than one instructor, classroom, grade level, school, setting, or time period. Sometimes results were reported separately for these different variables, and for different levels within the variables (e.g., for grades 6, 7, and 8 in schools A and B). Such presentations were more clearly indicative of a program's replicability than those in which the program was implemented across variables or levels but the data were reported in aggregated form only (e.g., reporting one combined mean for grades 6, 7, and 8). The JDRP approval rate was higher for submittals in which non-aggregated replication data were presented for at least one variable, than for those submittals reporting only aggregated replication data (65% vs. 49%, respectively).

Description of the Data Analysis Procedures

The methods used to analyze the data collected to substantiate claims of effectiveness ranged from descriptive statistics to complex multiple regression analyses. Besides descriptive statistics, the type of data analysis reported most frequently was the t-test (n = 147, or 63% of the submittals). Analysis of covariance (ANCOVA), analysis of variance (ANOVA), and nonparametric statistics were the next most frequently chosen analytical methods (see Table 11). All of the submittals reported the use of tests of statistical significance, with the exception of five submittals which provided descriptive statistics only. Of this latter group, only one was not approved by the JDRP.

Insert Table 11 about here

The adequacy of the procedures used to analyze the evaluation data was examined by identifying features which might negatively affect the believability and interpretability of the data. For 79 (or 34%) of the submittals reviewed, no problems in the data analyses were noted. In the other submittals, problems included the use of inappropriate or inadequate analysis procedures (n = 71, or 31%), omission of some relevant outcome data (n = 52, or 22%), and omission of information about the analysis procedures used, such as the p value or the name of the statistical test (n =

31, or 13%). The JDRP approval rate for submittals with no problems in data analysis was considerably higher than that for submittals having one or more problems: 71% vs. 50%, respectively.

Effect Sizes

The mean effect size over all the programs for which this statistic could be calculated was 0.89 (n = 263, SD = 1.10). Mean effect sizes were also calculated for the different levels of categorical variables descriptive of different features of the evaluation designs used. Table 12 presents these results, which are discussed here.

Insert Table 12 about here

Evaluator affiliation. The highest effect sizes were obtained by those programs evaluated by independent evaluators (M = 0.99), followed by programs evaluated by program staff (M = 0.91). Evaluator affiliation accounted for very little of the variance in the distribution of effect sizes.

Instrument type. Programs for which instruments were specifically developed to measure program outcomes had mean effect sizes almost twice as high as those programs relying on available published tests (M = 1.25 and M = 0.67, respectively). Of the evaluation characteristics examined for relationships to effect size, instrument type was the best single explanatory variable, accounting for 12% of the variance in the distribution of obtained effect sizes.

Design type. Designs based on randomized assignment of subjects to groups had higher mean effect sizes (M = 1.13) than those based on nonrandomized assignment (M = 0.92), and both of these had higher mean effect sizes than norm-referenced designs (M = 0.59).

Design quality. In this study, the higher the design quality, the higher the mean effect size that was found, ranging from 0.93 for high quality designs, to 0.89 for medium, and 0.67 for low. However, the strength of the relationship between design quality and effect size was not great (Pearson r = .09, n = 262), and design quality accounted for only a negligible proportion of the variance in the effect size distribution.

Data analysis quality. Programs for which only one or no flaws in the data analysis procedures were noted had higher mean effect sizes than those programs which had two or three problems

(M = 0.93 and 0.68, respectively). The extent of problems in the data analysis accounted for a very small proportion of the variance in effect sizes.

Effect size formulas. As was described earlier, adjustments based on suggestions from the meta-analytic literature were sometimes necessary when calculating effect sizes from the data reported in the JDRP submittals. When effect sizes were calculated from the basic effect size formula (treatment minus comparison group post means divided by the standard deviation of the comparison group), the mean effect size was considerably higher than when effect sizes were calculated using other formulas (M = 1.02 for the former, and M = 0.58 for the latter). The use of different effect size formulas accounted for 8% of the variance in the effect size distribution.

Combining Variables to Explain Effect Size

Multiple regression analyses were run on the 262 cases for which effect sizes had been computed. Effect size was specified as the dependent variable and type of objective, derivation of the outcome measure, type of evaluation design, evaluator affiliation, evaluation quality, quality of the data analysis, and formula for estimating effect size (basic vs. one of the adapted formulas) were the independent or explanatory variables. Dummy variables were created when categorical variables had more than two possible levels, resulting in a total of 13 explanatory variables being entered into the regression equation. A forward stepwise procedure was used, the results of which are presented in Table 13.

Insert Table 13 about here

The largest single contributor to the explanation of the variance in the effect size distribution was that the outcome measure was locally developed; the proportion of variance accounted for by this factor alone was 11.3%. The variable contributing the next largest amount to the proportion of explained variance (2.4%) was the type of effect size formula used (higher effect sizes were obtained with the basic formula than with its variations). Other variables which entered the regression equation were presence of an attitudinal objective, presence of a behavioral objective, and independent evaluator. Of the five variables which satisfied the criteria for entry into the regression equation, all were positively related to effect size except the presence of an attitudinal objective. The multiple R^2 resulting when these five independent variables were entered into the regression equation was 0.17.

Summary and Conclusions

The effect of the information explosion on the field of educational program evaluation has been to put evaluators in the embarrassing position, to paraphrase Glass (1976), of knowing less than we have proven. The hundreds of JDRP submittals which now exist are a perfect example of what Glass means. They describe educational programs and evaluation practice in a wide variety of content areas and from locations all across the United States. As such, they are a rich source of data about how to conduct and evaluate educational programs--a source that has gone largely untapped (with some exceptions: see Fang, 1981; Hamilton & Mitchell, 1979; Haney, 1978; The Network, 1978).

In this study, a total of 232 reports describing educational programs seeking JDRP approval during the years 1980 through 1983 were reviewed. The author developed a JDRP submittal analysis form to retrieve and document information about (among other things) educational program evaluation procedures actually being used in classrooms across the country. Found to be typical of evaluation practice, as represented by this group of programs, were these characteristics: (1) the evaluations were conducted by independent evaluators, either alone or in combination with program staff; (2) a single measuring instrument was used, usually a published test for which content validity and internal consistency were reported; (3) a single evaluation design was employed, most often being a nonequivalent control group design; (4) the quality of the evaluations (the degree to which threats to internal validity were controlled or eliminated) was not typically high enough to produce convincing evidence that the program produced the claimed effects; (5) evidence of replication of effects was gathered; (6) descriptive statistics and t-tests were used to analyze the data; (7) and the statistical significance of the obtained results was assessed.

Certain features of the evaluation procedures undertaken by the educational programs in order to demonstrate effectiveness were found more often in programs approved by the JDRP than in those not approved. These included: the presence of an independent evaluator affiliated with a research firm, the use of more than one evaluation design, the absence of obvious errors in the data analyses, and the implementation of evaluation designs of high quality (in this study, defined as elimination or control of threats to internal validity). This latter factor was found to be a particularly important consideration to the JDRP, with 79% of the evaluations rated as high or very high being implemented in programs which were approved by the JDRP, compared with only 32% of those rated as low or very low. This finding is consistent with the JDRP's stance that the ability of a program to demonstrate that observed effects can be attributed to program processes is of primary importance in their review process (Fang, 1981).

Of potentially greater importance to the improvement of educational program evaluation are the findings relating differences in evaluation characteristics to differences in the size of effects associated with the educational programs. While much has been written about the theoretical and statistical benefits and hazards of conducting various types of inquiries (see, for example, Boruch & McLaughlin, 1982; Campbell & Boruch, 1975; Gilbert, Light, & Mosteller, 1975; Kennedy, 1981; Rossi, 1979), it has been argued that what is really needed for the improvement of evaluation is publicly verified evidence of the usefulness of applications of evaluation methods in a variety of settings (Smith, 1979). Results of this study suggest that decisions about how to conduct evaluations or how to assess the meaning of program results should reflect an awareness that certain characteristics of evaluations may be differentially related to the size of effects detected. Illustrations of the data on which this conclusion is based follow.

In this study, the use of locally developed instruments was associated with higher effect sizes than the use of published tests. Because published tests are designed for maximum applicability across a wide range of educational experiences, they are more effective at measuring general achievements than specific learnings (Ball, 1981). As the match between the measuring instrument and specific program outcomes improves, other things being equal, the size of effects detected will increase. Because this is true, programs evaluated with the use of locally developed instruments may show larger effect sizes than those evaluated with published tests, even if the former programs are actually less effective than the latter.

Other features which were shown to be related to effect size were the type and quality of the evaluation design used, with higher effect sizes associated with randomized designs and designs of high quality. Recall that the magnitude of effect size is dependent on two factors: the difference between the treatment and comparison groups, and the amount of variance that exists within the study. To the extent that the evaluator can reduce extraneous variance through increased precision of measuring instruments, or through careful planning and implementing of the evaluation design, the size of the effects detected will increase, other things being equal (Hall, 1980; Sechrest & Yeaton, 1982). While these features are certainly desirable for all evaluation research, when they do not exist consistently across a sample of educational programs being compared, the interpretation of effect size for any given program is confounded with the quality of the evaluation design.

Finally, this results of this study suggest that there is a great deal of room for improvement in the quality of educational program evaluation being carried out in real-life classroom

settings. Fewer than half of the evaluations reviewed met the criteria for producing reasonably believable evidence of program effectiveness. This finding is particularly troublesome because programs which apply to the JDRP for validation represent some of the finest efforts being made in education in the United States today. Continued systematic study of these programs will contribute to our understanding of what makes educational programs effective, and continued systematic study of the benefits and hazards of applying different evaluation methods will help us improve the ways we go about assessing educational program outcomes.

Table 1

JDRP Decisions on Submittals by Year

Year	Submittals n	Approved n	(%)	Not approved n	(%)
1980	45	31	(69)	14	(31)
1981	68	43	(63)	25	(37)
1982	61	35	(57)	26	(43)
1983	58	35	(60)	23	(40)
Total	232	144	(62)	88	(38)

Table 2

Content Areas Addressed in JDRP Submittals

Content area	Programs n	(%)
Reading	86	(26)
Math	81	(25)
Special education	28	(09)
Career education	26	(08)
Language arts	22	(07)
Natural science	19	(06)
Social science	17	(05)
Health/physical education	15	(05)
Bilingual education	9	(03)
Gifted education	7	(02)
Vocational education	7	(02)
Writing education	6	(02)
Teacher education	4	(01)

(table continues)

Content area	Programs n	(%)
Arts/humanities	3	(01)
Migrant education	1	(01)
Other	14	(04)

Note. Based on N = 326 programs total. The sum of the frequencies exceeds 326, and the sum of the percentages exceeds 100 because some programs could be classified into more than one category.

Table 3

Number of Programs by Educational Level

	Programs	
Educational level	n	%
Preschool	12	4
K to grade 3	102	31
Grades 4 to 6	82	25
Grades 7 and 8	56	17
Grades 9 to 12	59	18
Post secondary	23	7

Note. Based on N = 326 programs total. The sum of the frequencies > 326 and the sum of the percentages > 100 because some programs spanned more than one of the above categories.

Table 4

Type of Educational Objectives Addressed in JDRP Submittals

Objective type	Submittals n	(%)
Cognitive only	157	(68)
Cognitive and behavioral	30	(13)
Cognitive and attitudinal	23	(10)
Cognitive, behavioral, and attitudinal	11	(05)
Behavioral only	4	(02)
Behavioral and attitudinal	2	(01)
Other	5	(02)
Total	232	(100)

Table 5

Evaluators' Affiliations

Types of evaluators	Submittals n	(%)[a]	
Program staff only	18	8	
Independent only	92	40	
Academic	57		46[b]
Research firm	28		22[b]
Staff plus independent	16	7	
"Other" only	27	12	
Combinations with "other"	23	10	
No information/cannot tell	56	24	

[a]Based on N = 232 submittals. [b]Based on n = a total of 125 evaluators identified as independent.

Table 6

Descriptions of Outcome Measures

Feature	Outcome measures n	% [a]	
Type			
Published	250	57	
Locally developed	129	29	
Modified	12	3	
Other	15	3	
Administration			
Norm-referenced	64	15	
At norming times		45	70 [b]
Not at norming times		19	30 [b]
Treatment/comparison groups	214	49	
At same times		212	99 [c]
Not at same times		2	1 [c]

(table continues)

Feature	Outcome measures n	% [a]
Validity information		
Face	16	4
Content	233	53
Construct	61	14
Criterion	33	8
Other	60	14
No information	99	23
Reliability information		
Stability	72	16
Equivalence	15	3
Internal consistency	166	38
Interrater	33	8
Other	104	24
No information	112	26

[a]Based on N = 438 instruments total. [b]Based on n = 64. [c]Based on n = 214.

Table 7

Evaluation Designs in the JDRP Submittals

Type of design	Submittals n	Submittals %
Nonrandomized untreated comparison pre-post	87	38
Nonrandomized untreated comparison post only	57	25
Nonrandomized alternate treatment pre-post	8	3
Nonrandomized alternate treatment post-only	5	2
Nonrandomized un- and alternate trt pre-post	5	2
Nonrandomized multiple time series	1	<1
Nonrandomized, other	2	1
National norms pre-post	61	26
National norms post only	21	9
State/local norms pre-post	6	3
Randomized untreated comparison pre-post	30	13
Randomized untreated comparison post only	9	4
Randomized alternate treatment pre-post	2	1
Randomized alternate treatment post only	1	<1
Randomized un- and alternate trt pre-post	4	2

(table continues)

Type of design	Submittals n	Submittals %
Randomized multiple time series	2	1
Randomized, other	2	1
One-group pre-post	33	14
One-group post only	8	3
One-group time series	5	2
Criterion-referenced	1	<1
One-group, other	8	3
Qualitative	1	<1
Cannot tell	4	2

Note. Percentages based on N = 232 submittals.

Table 8

Type Evaluation Design by JDRP Decision

Type of Design	Submittals n	(%)[a]	Approved n	(%)[b]	Not approved n	(%)[b]
Quasi-experimental	160	(73)	111	(65)	59	(35)
Norm-referenced	88	(38)	16	(67)	29	(33)
True experimental	50	(22)	33	(66)	17	(34)
One-group	50	(22)	35	(70)	15	(30)
Qualitative	1	(01)	1	(100)	0	(0)

Note. Total number of submittals = 232. The sum of the frequencies for type of design > 232 because many submittals reported more than one evaluation design.

[a]Based on N = 232 submittals. [b]Based on n for type of design.

Table 9

Quality of Evaluation Design by JDRP Decision

Quality rating	Designs n	(%)[a]	Approved n	(%)[b]	Not approved n	(%)[b]
Very high	20	(06)	15	(75)	5	(25)
High	139	(38)	111	(80)	28	(20)
Medium	135	(37)	93	(69)	42	(31)
Low	20	(06)	5	(25)	15	(75)
Very low	49	(13)	17	(35)	32	(65)

[a]Based on N = 363 total number of designs. [b]Based on n for rating category.

Table 10

Summary Data on Quality Ratings by Evaluation Design Type

		Quality rating			
		Very high/ High	Medium	Very low/ Low	
Type design	n	%	%	%	Mean rating
One-group	50	18	34	48	2.32
Norm-referenced	88	38	48	14	3.18
Quasi-experimental	170	47	37	17	3.22
True experimental	50	74	20	6	3.84
Qualitative	1	0	100	0	3.00
Total[a]	359	44	37	19	3.17

Note. Percentages based on row totals.

[a]Four designs coded as "cannot tell" were eliminated from this analysis.

Table 11

Data Analysis Methods in JDRP Submittals

Method	Submittals n	Submittals %
Descriptive statistics	232	100
T-tests	147	63
ANOVA	63	27
ANCOVA	70	30
Regression analyses	12	5
Nonparametric statistics	58	25
Qualitative analyses	1	<1

Note. The sum of the frequencies > 232 and the sum of the percentages > 100 because submittals could include more than one method.

Table 12

Mean Effect Size by Evaluation Design Characteristics

Characteristic	n	M	SD	R^2
Evaluator				.01
Independent only	137	.99	.91	
Staff only	18	.91	.67	
Combination	39	.77	.75	
Instrument type				.12
Published	135	.67	.53	
Locally developed	93	1.25	1.05	
Other	22	.80	.54	
Design type				.04
Norm-referenced	56	.59	.34	
Quasi-experimental	162	.92	.85	
Experimental	40	1.13	1.01	
Design quality				.01
Low/very low	32	.67	.67	
Medium	84	.89	.92	
High/very high	146	.93	.78	

(table continues)

Characteristic	n	M	SD	R^2
Data analysis problems				.02
0 or 1	215	.93	.85	
2 or 3	47	.68	.56	
Effect size formula				.08
Regular	187	1.02	.91	
Other	75	.58	.29	

Note. R^2 = the proportion of variance in the distribution of effect sizes accounted for when the evaluation design characteristic was considered the sole explanatory variable.

Table 13

Results of Stepwise Regression of Selected Variables on Effect Size

Step	Characteristic entered	R^2	Increase in R^2	Direction of influence
1	Local instrument	.113	.113	+
2	Basic effect size formula	.137	.024	+
3	Attitudinal objective	.152	.015	-
4	Behavioral objective	.163	.011	+
5	Independent evaluator	.173	.010	+

References

Ball, S. (1981). Outcomes, the size of the impacts, and program evaluation. New Directions for Program Evaluation, 9, 71-86.

Bernstein, I. N., & Freeman, H. E. (1975). Academic and entrepreneurial research. New York: Russell Sage Foundation.

Boruch, R. F., & Cordray, D. S. (1980). An appraisal of educational program evaluations: Federal, state, and local agencies (Contract No. 300-79-0467). Evanston, IL: Northwestern University.

Boruch, R. F., & McLaughlin, M. W. (1982). Boruch and McLaughlin debate mandated experimental methods. Evaluation News, 3(1), 11-20.

Bryant, F. B. (1982, October). Desegregation and black student achievement: Improving meta-analysis of quasi-experiments. Paper presented at the Joint Annual Evaluation Research Society/Evaluation Network Conference, Balitmore, MD.

Campbell, D. T., & Boruch, R. (1975). Making the case for randomized assignment to treatments by considering the alternatives: Six ways in which quasi-experimental evaluations in compensatory education tend to underestimate effects. In C. A. Bennett & A. A. Lumsdaine (Eds.), Evaluation and experiment (pp. 195-296). New York: Academic Press.

Campbell, D. T., & Stanley, J. C. (1963). Experimental and quasi-experimental designs for research. Chicago: Rand McNally.

Cook, T. D., & Campbell, D. T. (1979). Quasi-experimentation: Design and analysis issues for field settings. Chicago: Rand McNally.

Cook, T. D., & Gruder, C. L. (1978). Metaevaluation research. Evaluation Quarterly, 2, 5-51.

Datta, L. (1977). "There's more to effectiveness than achievement" and some other answers about the Joint Dissemination Review Panel. AERA/SIG Educational Research and Development Evaluators Newsletter.

Fang, W. L. (1981). The Joint Dissemination Review Panel: Can approved submittals be distinguished from rejected ones on the basis of presented evidence of effectiveness related to cognitive objectives? Unpublished doctoral dissertation, University of Virginia, Charlottesville.

Gilbert, J. P., Light, R. J., & Mosteller, F. (1975). Assessing social innovations: An empirical base for policy. In C. A. Bennett & A. A. Lumsdaine (Eds.), Evaluation and experiment (pp. 39-193). New York: Academic Press.

Glass, G. V. (1976). Primary, secondary and meta-analysis of research. Educational Researcher, 5(10), 3-8.

Glass, G. V. (1977). Integrating findings: The meta-analysis of research. Review of Research in Education, 5, 351-379.

Glass, G. V. (1980). Summarizing effect sizes. New Directions for Methodology of Social and Behavioral Sciences, 5, 13-31.

Glass, G. V., McGaw, B., & Smith, M. L. (1981). Meta-analysis in social research. Beverly Hills, CA: Sage.

Gordon, G., & Morse, E. V. (1975). Evaluation research. Annual Review of Sociology, 1, 339-361.

Hall, J. A. (1980). Gender differences in nonverbal communication skills. New Directions for Methodology of Social and Behavioral Sciences, 5, 63-77.

Hamilton, J. A., & Mitchell, A. M. (1979). Identifying and approving career education activities for national dissemination. The Vocational Guidance Quarterly, 28, 71-81.

Haney, W. (1978, March). The Follow Through planned variation experiment: Statistical conclusion validity. Paper presented at the Annual Meeting of the American Educational Research Association, Toronto, Canada.

Hirschi, T., & Selvin, H. C. (1967). Delinquency research: An appraisal of analytic methods. New York: Free Press.

Kennedy, M. M. (1981). The role of experiments in improving education. In C. B. Aslanian (Ed.), Improving educational evaluation methods: Impact on policy (pp. 67-77). Beverly Hills, CA: Sage.

Lipsey, M. L., Crosse, S., Dunkle, J., Pollard, J., & Stobart, G. (1985). Evaluation: The state of the art and the sorry state of the science. New Directions for Program Evaluation, 27, 7-28.

Lipsey, M. W. (1983, October). The wrong stuff: A methodological review of published evaluation studies. Paper presented at the Evaluation Research Society Annual Meeting, Chicago, IL.

McGaw, B., & Glass, G. V. (1980). Choice of the metric for effect size in meta-analysis. American Educational Research Journal, 17, 325-337.

McTavish, D. G., Brent, Jr., E. E., Cleary, J. D., & Knudsen, K. R. (1975). The systematic assessment and prediction of research methodology. Volume 1: Advisory report (Final Report on Grant OEO 005-P-20-2-74). Minneapolis, MN: Minnesota Systems Research, Inc.

Raizen, S. A., & Rossi, P. H. (Eds.) (1981). Program evaluation in education: When? How? To what ends? Washington, D. C.: National Academy Press.

Rossi, P. H. (1979). Past, present, and future prospects of evaluation research. In L. Datta & R. Perloff (Eds.), Improving evaluations (pp. 17-33). Beverly Hills, CA: Sage.

Sanders, J. R., & Nafziger, D. H. (1976). A basis for determining the adequacy of evaluation designs (Occasional Paper #6). Western Michigan University Evaluation Center, College of Education, Kalamazoo, MI.

Sechrest, L., & Yeaton, W. H. (1981). Assessing the effectiveness of social programs: Methodological and conceptual issues. New Directions for Program Evaluation, 9, 41-56.

Sechrest, L., & Yeaton, W. H. (1982). Magnitudes of experimental effects in social science research. Evaluation Review, 6, 579-600.

Smith, N. L. (1979). Requirements for a discipline of evaluation. Studies in Educational Evaluation, 5, 5-12.

Tallmadge, G. K. (1977). Ideabook: The Joint Dissemination Review Panel. Washington, D. C.: U. S. Office of Education.

Tallmadge, G. K., & Wood, C. T. (1978). User's guide: ESEA Title I evaluation and reporting system (rev. ed.). Mountain View, CA: RMC Research Corporation.

The Network. (1978). Program validation: Four case studies. Andover, MA: Author.

Worthen, B. R., & Sanders, J. R. (1973). Educational evaluation: Theory and practice. Belmont, CA: Wadsworth Publishing.